The ALICE Data Challenges

F. Rademakers[1]
For the ALICE Off-line Project – ALICE Collaboration

GSI, Darmstadt, Germany

Abstract

When fully operational the ALICE experiment will take data at a rate of 1.5 GB/s. This rate is more than an order of magnitude higher than that of the other LHC experiments. To be ready for this very high rate and to understand early on what the possible problems might be, we have decided to start the *ALICE Data Challenges* (ADC). The idea is that the ADC's will be repeated once or twice per year till LHC goes online. The unique aspect of the ADC's is that it involves the complete ALICE data taking chain: from DAQ, via event builder, via filter, to object database, to mass storage system. Progressively, with each ADC we'll try to achieve higher date rates, more realistic event filters and raw data models.

Keywords: ALICE, CHEP, OO, data challenge, ROOT, MSS, DAQ

1 Introduction

ALICE [1], when running at full performance, will produce data at a rate of 1.5 GB/s. This rate is an order of magnitude higher than that of any of the other LHC experiments. To make sure that the proposed CERN infrastructure for Central Data Recording (CDR) and Mass Storage Systems (MSS) takes ALICE's special requirements into account, and to be confident that the rest of the ALICE data chain is up to the task we decided to start the ALICE Data Challenges (ADC). The goal of these ADC's is to test all critical components on the path from data acquisition (DAQ) to MSS. With each successive ADC we try to achieve higher data rates, use more sophisticated data models and try new technologies (networking, CPU's, MSS's, etc.). The ADC's must give us confidence that we can handle the 1.5 GB/s in 2005.

2 The First ADC

We started planning the first ADC (ADC1) end of January 1999. The fairly modest goal was to be able to store within 5 days 10 TB of "raw" data, generated by the DATE [4] system, in a ROOT [2] database on HPSS [3] (i.e. a rate of 25 MB/s). With ADC1 we would test the following technologies:

- DATE
- ROOT
- HPSS
- PC's
- Linux
- Gigabit Ethernet

Where DATE is the ALICE DAQ system, ROOT the Object Oriented data handling framework and HPSS a commercial mass storage system.

We prepared ADC1 in close collabaration with the CERN IT/PDP group who is in charge of the HPSS infrastructure. It turned out there was only one time slot in which we could have dedicated access to the HPSS system and the necessary Gigabit Ethernet segments. This was just before the CERN experimental program would restart in late March, which gave us a very thight schedule to setup the test environement and build the necessary software components.

2.1 ADC1 Setup

As data source for ADC1 we used the DATE setup of the NA57 experiment in the CERN North Area (Prevessin). This setup consisted of 9 rack mounted PowerPC's, running the *local data concentrator* (LDC) component of the DATE system, which generated bit patterns at a rate of 5 MB/s per machine. The data was then sent over a Gigabit Ethernet link from the North Area to a 5 node Linux PC cluster located at the CERN main computer center. On each PC there was one DATE *global data concentrator* (GDC) running. Each GDC was receiving data from two LDC's (except for one node where the GDC did only handle one LDC). The GDC's combined the two LDC streams and piped the resulting "event" into the ROOT based *objectifier* program. The objectifier created a simple event object containing the raw event bit pattern (400-500 KB) and some book keeping data (date, time, run number, event number, etc.). The event objects were then stored into a ROOT database. In addition to the raw event database the objectifier created two additional databases: a *tag* database, containing some summary information for each event, and a *run catalog* database, containing information on each raw event database file. Whenever the file size of a raw event database reached a certain limit (we varied this limit from 500 MB - 1.5 GB) the file was closed and a new one created. In the background there was a server running that copied the closed raw event databases via the RFIO [5] `rfcp` command to the HPSS servers. After a file was copied it was deleted. The net output rate via `rfcp` was 5 MB/s per node. Figure 1 shows the ADC1 setup.

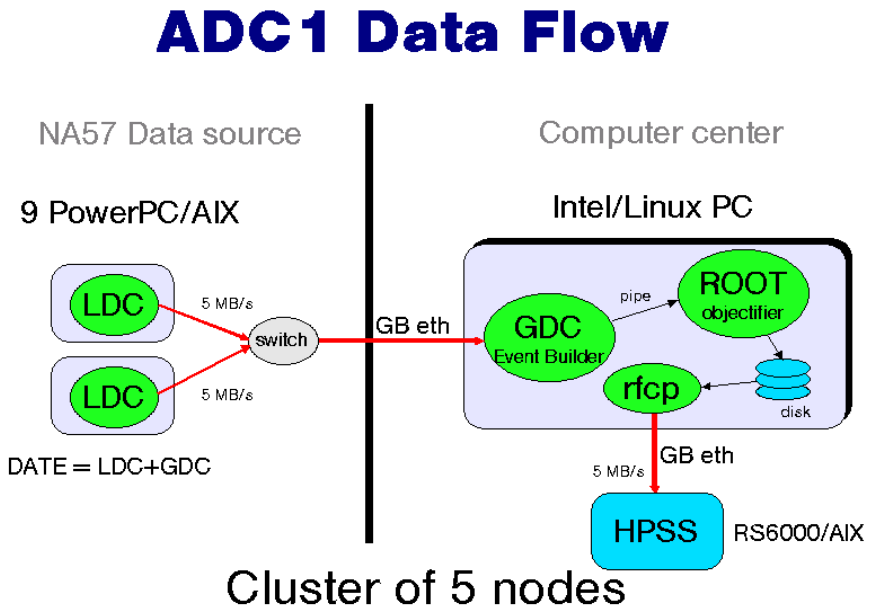

Figure 1: The ADC1 setup

2.2 ADC1 Results

ADC1 started perfectly and after one day we had stored into HPSS more than 2 TB of data. This was too good to be true, and indeed during the second day the HPSS servers, 2 IBM AIX machines, went down. It turned out that the I/O backplane containing the Gigabit Ethernet cards had physically "melted down" due to the continuous high load. Trying to identify and diagnose the problem caused ADC1 be interrupted or slowed down for about 2.5 days. IBM could not deliver the needed spare parts in time, so we had to decide to stop ADC1 or continue without HPSS. We decided for the latter and set the system up so that the files were still copied to the HPSS front-end servers, but instead of moving the files into HPSS they were directly deleted.

We also experimented with the ROOT "on the fly" compression feature to see if it was feasible to write a compressed raw event database. Although the raw event objects could be compressed by a factor 3, the amount of CPU needed was too high. Using compression we could only write 500 KB/s (i.e. 1.5 MB/s uncompressed) instead of 5 MB/s.

After 7 days (two longer than planned) we stopped ADC1. The "Grand Total" after 7 days of running is shown in Table I.

The main conclusions of ADC1 are:

- A HPSS hardware weak point has been discovered and fixed.
- We can objectify a simple ALICE raw data stream of 1.5 GB/s. With a farm of 300 PC's we could already do it now. Compression should be possible in 2005.
- DATE and ROOT are, simple, well understood and seem a viable solution.
- The system is totally scalable since there is no single contention point (like, a lock manager, etc.).

Number of files	15436
Number of events	16229520
MB's in	7261382
MB's out	6896198
Aggregate rate in (MB/s)	14.7
Aggregate rate out (MB/s)	13.9
Total moved to HPSS	6.9 TB

Table I: Grand total for ADC1 after exactly 7 days of running

3 The Second ADC

At the moment we are setting up ADC2 which should be run begining of March 2000. For ADC2 we have raised the stakes considerably. The goals are:

- Store data with a rate of 100 MB/s for 10 days
- Store data in HPSS and CASTOR [6] (5 days in HPSS, 5 days in CASTOR)
- Use first version of ALICE raw data format
- Use "real" simulated ALICE event as input instead of random bit patterns
- Try to do simple online filtering (finding of two hard tracks)

To be able to reach 100 MB/s we need to increase the number of data sources compared to ADC1. For that we will use a second batch a DATE LDC's running on a number of Intel PC's in the ALICE DAQ lab. Also by trying to avoid the intermediate step of storing the raw event database on the local disk of each cluster node we expect to be able to increase the throughput

per cluster node. However, this would be offset again by adding an online filter step (μ Level 3). Therefore, we expect we will need a cluster of at least 20 nodes to achieve 100 MB/s.

CASTOR is an alternative MSS currently under development at CERN. In its first incarnation it is an extension of the RFIO system while later more specific MSS components will be added. In ADC2 we decided to beta-test CASTOR before it will be taken in production by the COMPASS experiment later this year. See figure 2 for the planned ADC2 setup. If there are no major breakdowns we should be able to record about 86 TB during ADC2.

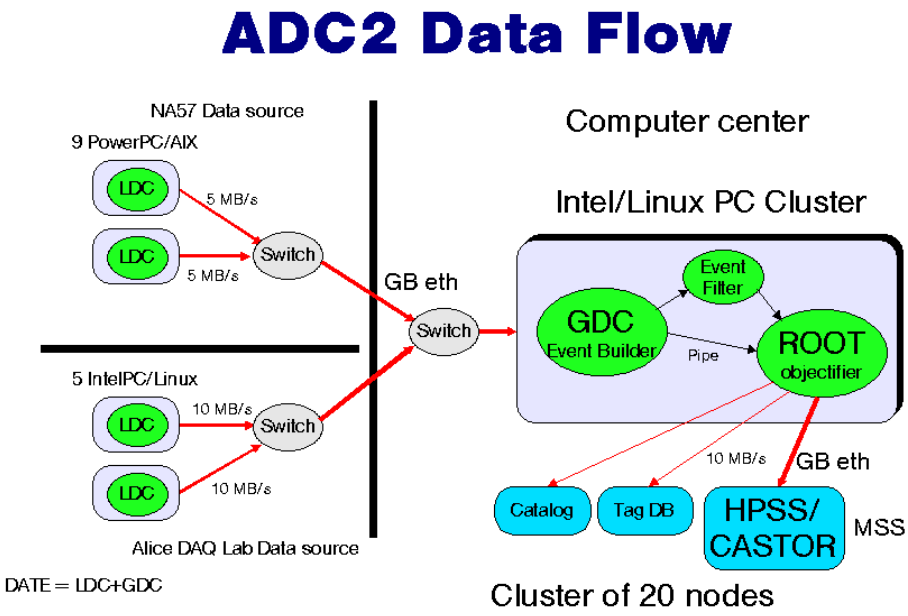

Figure 2: The planned ADC2 setup

4 Conclusion

The first ADC proved to be a very useful excercise. It allowed the on-line and off-line groups to work together and integrate their technologies in an early stage (6 years before experiment startup).

The ADC's allow us to track different technologies in a controlled environment, giving a better understanding of the possible weak and strong points in the critical data taking and recording chain. In future ADC's we will continue to increase the data rates, introduce more sophisticated event models, try different level-3 filtering techniques and test other MSS systems. In addition we plan to extend the ADC's to include remote data access strategies (simulating regional centers).

5 Acknowledgements

We would like to acknowledge the outstanding participation of the CERN IT/PDP group in the organization and running of the ADC's.

References

1 ALICE Technical Proposal for A Large Ion Collider Experiment at the CERN LHC, CERN/LHCC/95-71, 15 December, 1995.

2 R. Brun and F. Rademakers, ROOT – An Object Oriented Data Analysis Framework, Proceedings AIHENP'96 Workshop, Lausanne, September 1996, Nucl. Inst. & Meth. in Phys. Res. A 389 (1997) 81-86. See also http://root.cern.ch.
3 http://www.sdsc.edu/hpss/hpss.html.
4 ALICE DATE V3.5 User's guide, ALICE internal note INT-99-46, December 1999.
5 http://wwwinfo.cern.ch/pdp/serv/shift.html.
6 http://wwwinfo.cern.ch/pdp/castor/.